What is a Data Warehouse?

- **Definition**: A Data Warehouse (DW) is a **subject-oriented, integrated, time-variant, and non-volatile** collection of data designed to support decision-making processes. This definition is given by **W. H. Inmon**.
- **Purpose**: Provides a consolidated, historical data platform to aid in decision support and analysis.

Key Characteristics of Data Warehouses

1. **Subject-Oriented**:
   - Organized by major subjects like customer, sales, or product.
   - Focuses on decision-making, excluding operational or transactional data.
2. **Integrated**:
   - Combines data from multiple, heterogeneous sources (e.g., relational databases, flat files).
   - Data cleaning and transformation ensure consistency in naming conventions, units, and formats.
3. **Time-Variant**:
   - Stores historical data spanning 5–10 years or more for trend analysis.
   - Unlike operational databases, it contains time elements for historical tracking.
4. **Non-Volatile**:
   - Once data is stored, it remains unchanged. Updates or deletions do not occur in the warehouse.
   - Involves **initial loading** and **data access** but excludes transactional processes.

OLTP vs. OLAP

- **OLTP (Online Transaction Processing)**:
  - Used for daily operations.
  - Current, detailed data in an application-oriented structure.
  - Supports many users with fast, repetitive transactions.
- **OLAP (Online Analytical Processing)**:
  - Designed for decision support and complex queries.
  - Historical, summarized data in a subject-oriented structure.
  - Serves fewer users and handles large data queries.

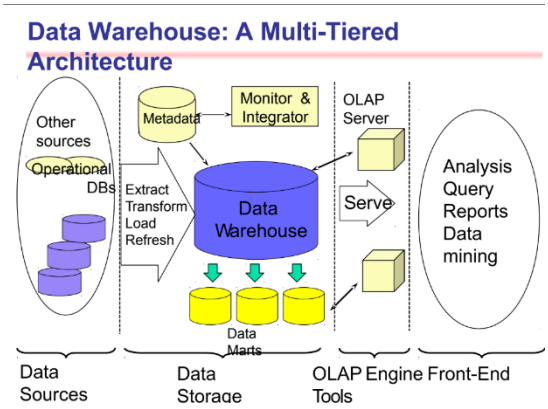
Why a Separate Data Warehouse?

- **Performance**: Tailored for specific needs—OLTP for transactions, OLAP for analysis.
- **Functionality**: OLAP requires historical data, aggregation, and reconciliation of inconsistent data sources.
- **Data Consolidation**: Integrates and transforms data from various sources.

Metadata Repository Summary:

- **Definition:** Metadata describes the structure and objects in a data warehouse.
- **Key Elements:**
  - **Structural Metadata:** Includes warehouse schema, views, dimensions, hierarchies, derived data definitions, and data mart details.
  - **Operational Metadata:** Tracks data status (active, archived, purged), usage statistics, error reports, audit trails, and performance metrics.
  - **Summarization Details:** Covers algorithms used for data aggregation and mapping from operational systems to the warehouse.
  - **Business Metadata:** Includes business terms, definitions, and data ownership details.

Data Warehouse Architecture



Multi-Tiered Architecture of a Data Warehouse, which is commonly used to manage, store, and analyze large volumes of data. Here's an explanation of its components and workflow:

1. Data Sources

This is the foundation where raw data originates, including:

- **Operational Databases**: These are transactional systems such as ERP (Enterprise Resource Planning) and CRM (Customer Relationship Management).
- **Other Sources**: Includes flat files, spreadsheets, or external systems such as third-party APIs or web logs.

---

2. ETL Process (Extract, Transform, Load, and Refresh)

This layer is responsible for **data integration** and preparing the data for storage:

- **Extract**: Pulls data from multiple sources.
- **Transform**: Cleans and processes the data to ensure consistency (e.g., resolving naming conflicts, unit conversions).
- **Load**: Stores the transformed data into the data warehouse.
- **Refresh**: Periodic updates to keep the data warehouse current.

---

3. Data Storage

- **Data Warehouse**: Serves as the central repository of integrated and historical data. It contains cleaned and consolidated data for analysis.
- **Metadata**: Provides information about the data, such as its source, transformation rules, and lineage.
- **Monitor & Integrator**: Tracks data refresh rates, ensures consistency, and monitors the overall health of the warehouse.
- **Data Marts**: Subsets of the data warehouse tailored to specific departments or business functions (e.g., Sales Data Mart or HR Data Mart).

4. OLAP (Online Analytical Processing) Server

- **Purpose**: Supports complex queries and multidimensional analysis (e.g., slicing, dicing, roll-up, drill-down).
- **Serve**: Processes and delivers the data to front-end tools for reporting or analysis.

---

5. Front-End Tools

These are the interfaces for users to interact with the data warehouse:

- **Analysis**: Includes tools for statistical and predictive analysis.
- **Query**: Ad hoc querying tools for exploring data.
- **Reports**: Predefined and customized reports for decision-making.
- **Data Mining**: Tools for discovering patterns, trends, and actionable insights from the data.

Workflow Summary

1. Data is collected from various **sources** and passed through the **ETL process**.
2. Cleaned and processed data is stored in the **Data Warehouse**.
3. Subsets of data are created in **Data Marts** for specific use cases.
4. The **OLAP Server** facilitates analytical queries and processing.
5. Finally, **front-end tools** enable users to visualize, query, and analyze the data for insights.

Three Data Warehouse Models

1. Enterprise Data Warehouse (EDW)

- **Definition**: A centralized repository that collects, stores, and manages data about all subjects across the entire organization.
- **Key Features**:
  - Comprehensive and integrates data from all operational databases and external sources.
  - Designed to support enterprise-wide decision-making.
  - Typically large-scale and handles data on a global or organizational level.
- **Use Case**: A retail company uses an EDW to analyze sales, inventory, customer behavior, and supply chain performance across all stores globally.

---

2. Data Mart

- **Definition**: A smaller, more focused subset of the enterprise data warehouse, tailored to meet the needs of specific business functions or groups.
- **Types**:
  - **Independent Data Mart**:
    - Created directly from external sources or operational databases.
    - Operates independently without relying on an enterprise data warehouse.
    - Example: A marketing team builds a data mart to analyze customer surveys and third-party market research data.
  - **Dependent Data Mart**:
    - Extracted from the enterprise data warehouse.
    - Offers consistency since it inherits its data from the central EDW.
    - Example: A finance department creates a dependent data mart for detailed budget tracking.
- **Scope**: Limited to a particular department or function (e.g., Marketing, Sales, or HR).

---

3. Virtual Data Warehouse

- **Definition**: A logical model that provides a set of views over operational databases without physically storing data in a central location.
- **Key Features**:
  - Data is not integrated or stored in one location but accessed dynamically from various sources.
  - Only some **summary views** are materialized (pre-computed for quick access), while others are generated on-the-fly.
  - Relies heavily on operational systems for up-to-date information.
- **Advantages**:
  - Cost-effective and requires minimal storage space.
  - Quick implementation compared to an EDW.
- **Limitations**:
  - Performance can be slower due to real-time access to source systems.
  - Limited ability to handle complex queries compared to an EDW.
- **Use Case**: A small business uses a virtual warehouse to analyze sales trends by directly accessing operational databases and external online sales platforms.

Schemas in Data Warehousing

1. **Star Schema**:
    - Central fact table connected to dimension tables.
    - Simple and efficient for querying.
2. **Snowflake Schema**:
    - Refinement of the star schema with normalized dimensions.
3. **Fact Constellation (Galaxy Schema)**:
    - Multiple fact tables sharing dimension tables.

OLAP Operations

1. **Roll-Up**: Summarize data by climbing hierarchies or reducing dimensions.
2. **Drill-Down**: Move from summarized data to detailed levels.
3. **Slice and Dice**: Project and select subsets of data.
4. **Pivot**: Rotate data for visualization from different perspectives.
5. **Drill-Through**: Access detailed data in the operational system.

Design and Implementation

- **Top-Down vs. Bottom-Up**:
    - Top-Down: Start with overall planning and design.
    - Bottom-Up: Start with experiments and prototypes.
- **Typical Process**:

1. Choose a business process (e.g., sales).
2. Define the granularity (level of detail).
3. Identify dimensions and measures.

Usage of Data Warehouses

1. **Information Processing**: Querying, reports, and statistical analysis.
2. **Analytical Processing**: OLAP operations for multidimensional analysis.
3. **Data Mining**: Discover patterns and build models for prediction.

Efficient Computation of Data Cubes

- **Cube Computation**:
    - Extends SQL for aggregating data across dimensions.
    - Challenges: Storage and processing of large, multi-dimensional datasets.